The Interestingness of Images

Michael Gygli[1,2] Helmut Grabner[1,2] Hayko Riemenschneider[1]
Fabian Nater[2] Luc Van Gool[1,2,3]

[1]Computer Vision Laboratory, ETH Zurich
[2]upicto GmbH, Zurich
[3]ESAT - PSI / IBBT, K.U. Leuven
{gygli, grabner, hayko, vangool}@vision.ee.ethz.ch {gygli, grabner, nater}@upicto.com

Abstract

We investigate human interest in photos. Based on our own and others' psychological experiments, we identify various cues for "interestingness", namely aesthetics, unusualness and general preferences. For the ranking of retrieved images, interestingness is more appropriate than cues proposed earlier. Interestingness is, for example, correlated with what people believe they will remember. This is opposed to actual memorability, which is uncorrelated to both of them. We introduce a set of features computationally capturing the three main aspects of visual interestingness that we propose and build an interestingness predictor from them. Its performance is shown on three datasets with varying context, reflecting diverse levels of prior knowledge of the viewers.

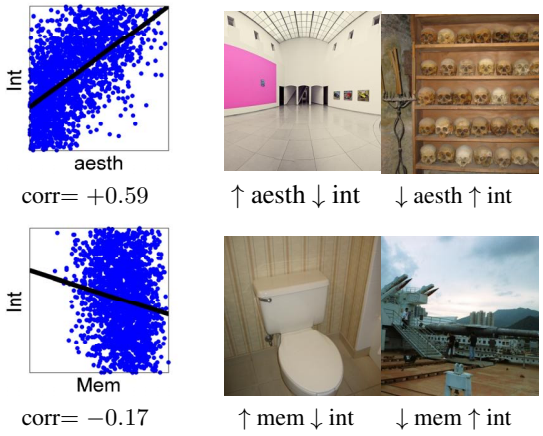


Figure 1: Interestingness compared to aesthetics and memorability.

1. Introduction

With content-based image retrieval on the rise, there is a parallel increase in the study of cues that could help in ranking the retrieved images. These include image quality [15], memorability [14] and aesthetics [10, 11]. Yet, a measure that would seem more relevant to automatically quantify is how interesting people find an image and this "interestingness" has hardly been studied so far. Apart from retrieval, other applications like video summarization or automated camera hand-over are also bound to benefit.

The most related work might be that by Dhar *et al.* [11], who used their high-level aesthetics features to train a classifier on Flickr's interestingness. It can predict the interestingness of these images well, but it is questionable that these results can be generalized to other image datasets. Flickr's interestingness [7] is based on social behavior, *i.e.* according to the uploader's reputation and a non-disclosed ratio between views, favorites and comments on the images. This measure has not been shown to relate to what people find interesting in images. For example, images where interest is caused through negative emotions (disgust, disturbance, threat, etc.) tend to get low Flickr interestingness. Users will hardly select such images as a favorite.

In our own series of psychological experiments we analyze "interestingness" and how it relates to measures such as aesthetics and memorability (Fig. 1). There exists indeed a strong correlation between aesthetics and interestingness (Fig. 1, top row). However, what is interesting does not necessarily need to be aesthetically pleasing, *e.g.* the image of skulls is interesting, even though it is not aesthetic. While one would also expect a high correlation of memorability and interestingness, our experiments indicate the contrary (Fig. 1, bottom row). More details are to follow.

In this paper we (i) investigate what arouses human interest (Sec. 2) and show that it is fundamentally different from other properties such as memorability; (ii) propose a set of features able to computationally capture the most important aspects of interestingness (Sec. 3); (iii) show the performance of these features and an interestingness predictor built from them, on three datasets with varying levels of context (Sec. 4); (iv) show that the context within which an image is viewed is crucial for the appraisal of interestingness.

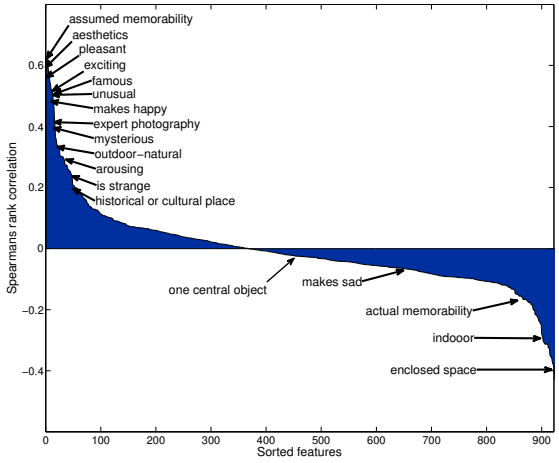
2. What causes human interest?

In his seminal work Berlyne [2] introduced four variables affecting interest: *novelty*, *uncertainty*, *conflict* and *complexity*. He showed that new, complex and unexpected events are a strong trigger of interest. Recent psychological research extends Berlyne's theory, *e.g.* Silvia [20] who analyzes the effects of complexity and understandability on interest. The more computational approach in [19] concurs with these ideas. Biederman and Vessel [3] explain interest with perceptual pleasure, resulting from comprehensible information and newly activated synapses. They furthermore found that natural scenes with wide landscapes are preferred over man-made scenes. Other cognitive work by Chen *et al.* [9] identifies novelty, challenge, instant enjoyment, and demand for attention as sources of interestingness. While Smith and Ellsworth [21] found that high pleasantness is a major aspect of interestingness, recent studies [24] indicate otherwise for images with polygons and paintings.

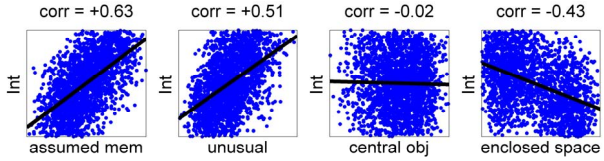
Given the lack of clear-cut and quantifiable psychological findings, we investigate the correlation of interestingness with an extensive list of image attributes, including emotional, aesthetic and content related aspects. We use the dataset of Isola *et al.* [14], extended in [13] to explore memorability. In Fig. 2a we relate the provided image attributes to the interestingness ground truth we collected (*c.f.* Sec. 4.3). This figure shows the Spearman rank correlation of all attributes and highlights several with high correlations (either positive or negative). Fig. 2b shows the correlations of four example attributes in more detail. In keeping with the work in psychology we find three main groups with high influence: *novelty/unusualness* (attributes: unusual, is strange, mysterious), *aesthetics* (attributes: is aesthetic, pleasant, expert photography) and *general preferences* for certain scene types (attributes: outdoor-natural vs. indoor and enclosed spaces).

As the predictability of related concepts (*e.g.* aesthetics) has been approached successfully in the past, there is good hope that we can computationally predict interestingness, based on the above cues. This assumption is supported by our experiments. When comparing the data of [13] with our own we find that people agree, to a large extent, on which images are interesting, despite personal preferences (*c.f.* Sec. 4.3). This observation of a high inter-individual agreement for real-world images was also shown by [26].

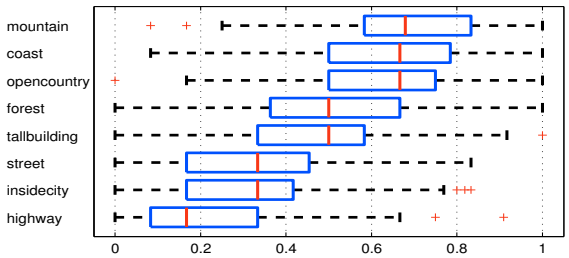
The cues that we implemented were selected on the basis of their experimentally verified correlation with interestingness. That is important, as intuition can often be misleading. For instance, Isola *et al.* [13] have shown that human prediction of what is memorable (*i.e.* assumed memorability) is negatively correlated with actual memorability. Interestingness, on the other hand, has its highest correlation with this assumed memorability. What a human observer finds interesting is what he wants to remember and believes he will. Unfortunately the latter is often not the case.



(a) Interestingness correlated with an extensive set of image attributes, based on the data of [13]. We compare the attributes to our interestingness score, collected as described in Sec. 4.3.



(b) Correlations of noteworthy attributes from above and interestingness.



(c) Correlation of scene categories and interest on the dataset of [18], interestingness scores obtained as described in Sec. 4.2.

Figure 2: What aspects relate to interestingness?

Additionally, we investigated the preference for certain scene types (Fig. 2c) and found, in agreement with [3], that people prefer natural outdoor scenes rather than man-made scenes. While interestingness is higher for images containing sky, actual memorability decreases if sky is present. Indeed, when comparing actual memorability and interestingness, we find them to be negatively correlated[1]. Nevertheless, we believe that also for applications like selecting images for advertisements (mentioned by Isola *et al.* [13]), it makes more sense to select an interesting image, than a memorable but dull one. In the end, the goal is not to have people remember the image, but the message combined with a positive connotation.

[1]The complete experimental data is available on the authors' webpage

3. Computational approach for interestingness prediction

In this section we propose features that computationally capture the aspects/cues of interestingness which we found most important (*c.f*. Fig. 2) and are implementable: unusualness, aesthetics and general preferences. Then, we use these to predict the interestingness of images. Thereby we build upon the set of features we used in our previous work for image sequences [12] and extend it with additional features suitable for single images.

Formally, given an image I we are looking for an interestingness score s. Our pipeline to achieve this task consists of two stages: (i) exploring various features to capture each of the above cues for interestingness and (ii) combining these individual features.

3.1. Unusualness

As said, unusualness/novelty is an important cue for interestingness. Abnormality detection goes in that direction, but most contributions consider constraint surveillance settings with fixed cameras. Instead, we want to capture unusualness in single images from arbitrary scenes. We propose two different methods, one relying on global image descriptors and one on image parts.

Global outliers. We detect global outliers in the dataset by applying the Local Outlier Factor (LOF) algorithm [6] to global image descriptors. LOF gives a measure to what degree a data point (an image, in our case) is outlying, taking into account its k nearest neighbors. It is called local, as the outlier factor is calculated with respect to the density of its closest cluster. In all our experiments we use a 10-distance neighborhood and as features (i) the raw RGB pixel values $s^{unusual}_{pixel}$, (ii) GIST [18] $s^{unusual}_{gist}$ and (iii) Spatial Pyramids on SIFT histograms [16] $s^{unusual}_{pyr}$. We use these features, as they have been found effective for scene categorization [18, 16] and image classification [16].

A similar idea was used by Datta and Wand [10], where they propose a *Familiarity* feature. This measure is computed as the average distance of a test image to the *k* closest training images (based on local features). The higher this distance the less familiar (more unusual) an image. Interestingly they found this feature to play a crucial role in the classification of image aesthetics (*c.f*. Fig. 1, correlation of interestingness and aesthetics $\rho = 0.59$).

Composition of parts. We also propose an approach that operates on image parts. It is inspired by the work of Boiman and Irani for the detection irregularities in images and videos [4]. Ensembles of patches from the tested image are matched against patches from the database, which defines what is "regular". If there is a good match, it is considered regular, otherwise as irregular. Instead of using square patches, overlapping between foreground and background, we over-segment the image using superpixels (SLIC [1]). This allows for a delineation of the image parts. We model the image as a graph with superpixels as nodes. The graph's energy determines how unusual the configuration of patches is:

$$E(\mathbf{L}) = \sum_{i \in \mathcal{S}} D_i(l_i) + \lambda \sum_{\{i,j\} \in \mathcal{N}} V(l_i, l_j) \qquad (1)$$

where $\mathcal{S}$ is the set of superpixels and $\mathcal{N}$ the set of superpixel neighbors. $D_i(l_i)$ denotes the unary cost of assigning label l to the superpixel i. The label space $\mathcal{L}$ is the set of images in the database (*i.e*. $|\mathcal{L}|$ is equal to the number of database images). The unary cost $D_i(l_i)$ is the Euclidean distance in the descriptor space of a superpixel i to the nearest-neighboring superpixel in the database with label l. The set of descriptors is that of [22], which includes features such as SIFT, Texton and Color histograms as well as location information. The binary terms $V(l_i, l_j)$ denote the cost of two neighboring nodes taking labels l_i and l_j, respectively. They encourage label smoothness, as $V(l_i, l_j) = 0$ if $l_i = l_j$ and 1 otherwise, *i.e*. a simple Potts penalty. Empirically we found the weighting parameter λ to be robust and we keep it fixed at 0.02.

To find the labeling that minimizes the cost, we apply MAP inference based on a standard GraphCut algorithm [5]. With $\mathbf{L}$ being that optimal labeling, the *unusualness by composition* is defined as $s^{unusual}_{compose} := E(\mathbf{L})/|\mathcal{S}|$, *i.e*. the energy in the Markov Random Field normalized by the number of superpixels.

Intuitively this feature encodes how well an image can be composed with parts of images in the database, while encouraging the composition of connected regions from only one image.

3.2. Aesthetics

To capture the aesthetics of an image, we propose several features that are rather simple in comparison to other, more extensive works in the area. For example [11] uses content preferences, such as the presence of people and animals or the preference for certain scene types to classify aesthetically pleasing images. We capture such general preferences with global scene descriptors in Sec. 3.3. For predicting aesthetics, we focus on capturing visually pleasing images, without semantic interpretation.

Colorfulness. We measure colorfulness as proposed by Datta and Wang [10], *i.e*. as the Earth Mover distance (in the LUV color space) of the color histogram of an image H_I to a uniform color histogram H_{uni}. A uniform color histogram is the most colorful possible, thus the smaller the distance the more colorful the image, $s^{aesth}_{colorful} := -\mathrm{EMD}(H_I, H_{uni})$.

Arousal. Machadjik and Hanbury [17] extracted emotion scores from raw pixels. Their features are based on the

empirical findings of [25], which characterized emotions that are caused by color using the space of *arousal*, *pleasure* and *dominance*. As interest is correlated with arousal (*c.f.* Fig. 2a), we use an arousal score as in [17]. It is calculated as the average over all pixels p of the image as $s^{aesth}_{arousal} := \sum_p -0.31 \text{ brightness}(p) + 0.60 \text{ saturation}(p)$.

Complexity. To capture the complexity of an image, we compare its size after JPEG compression against its uncompressed size, *i.e.* $s^{aesth}_{complex} := \frac{\text{bytes}(\text{compress}(I))}{\text{bytes}(I)}$. We use JPEG as it is a lossy compression, which compresses an image according to the human visual system [27]. If the compression rate is high $s^{aesth}_{complex}$ is low, as there is little visually important information in the image.

Contrast. We use the same contrast quality measure as [15], *i.e.* we calculate the minimal range of the 98% mass of a gray-scale histogram to obtain $s^{aesth}_{contrast}$.

Edge distribution. Following [15] we calculate the image bounding box that contains 98% of the edges in each dimension (*i.e.* along x and y). Smaller bounding boxes typically correspond to less clutter and a more uniform background. We use $s^{aesth}_{edges} := 1 - w_x w_y$, with w_x and w_y being the box's normalized width and height;

3.3. General preferences

Following the observation that certain scene types tend to be more interesting than others (*c.f.* Fig. 2c and [3]), we propose to learn such features from global image descriptors. We train a Support Vector Regressor (ν-SVR [8]) on raw RGB-pixels s^{pref}_{pixel}, GIST [18] s^{pref}_{gist}, spatial pyramids of SIFT histograms [16] s^{pref}_{pyr}, and color histograms s^{pref}_{hist}. Spatial pyramids and GIST are known to capture scene categories well. RBF kernels served our purposes in all cases. We used $\nu = 0.5$ and optimized the parameters γ and C through grid search on the validation set. We tested $C = \{2^{-5}, 2^{-3}, ..., 2^7, 2^9\}$ and $\gamma = \{2^{-15}, 2^{-13}, ..., 2^1, 2^3, 2^5\}$.

3.4. Combination

The scores obtained from the respective features are first normalized with respect to their mean and variance. Second, they are mapped into the interval $[0, 1]$ using a sigmoid function $\bar{s} = \frac{\exp(\mu s)}{1+\exp(\mu s)}$ where the parameter μ is estimated using least-square minimization on the validation set. To combine the individual features, we perform greedy forward feature selection. Starting from the best single feature, we select additional features until the combination does not improve further, as a quality measure using Spearman's ρ. As a model we use a simple linear combination $\bar{\mathbf{s}}_{comb} = \mathbf{w}^{\mathrm{T}} \bar{\mathbf{s}}_{sel}$, where $\bar{\mathbf{s}}_{sel}$ is the vector of selected features. The weights are trained using least-squares.

As we use a linear model that assumes uncorrelated features, we also applied whitening to decorrelate the features before training the model. We define $\bar{\mathbf{s}}_{decorr} = \mathbf{\Sigma}^{-1/2} \bar{\mathbf{s}}$, where $\mathbf{\Sigma}$ is calculated on the training set. This whitening step leads to only a marginal improvement, suggesting that the features are indeed complementary (*c.f.* Tab. 1).

4. Experiments

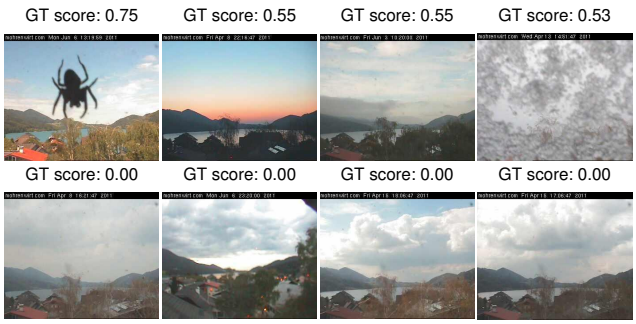
In this section we discuss the performance of the different interestingness features. As we will see, the strength of the contextual cues that are relevant in the tested setting determines – in part – which types of features are most effective in capturing interestingness. First, we specify the selection of parameters and the evaluation criteria. Then, we run through the results for three datasets.

Parameters. For the features based on raw pixels ($s^{unusual}_{pixel}$ and s^{pref}_{pixel}) we used downscaled images of 32×32 pixels, which is the size we found to work best. This agrees with [23], where it was shown sufficient to capture scene types and important objects. For each dataset we use a training/validation/test split. The the training set serves as a database for all the outlier methods, *i.e.* their response is high if a test image is not somehow similar to the training data. As for the general preference features, we trained the ν-SVR on the training set and optimized the hyperparameters using grid search on the validation set. The estimation of μ for the sigmoid function of each feature and the feature selection and estimation of the weight vector $\mathbf{w}$ for the combinations are also performed on the validation set. Both, the test and validation set consist of 240 randomly selected images (unless specified otherwise).

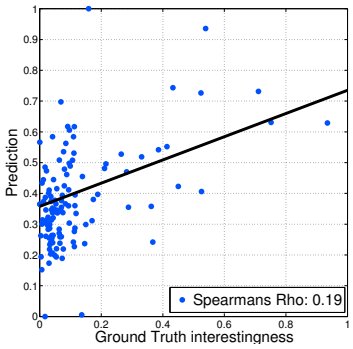
Evaluation. In order to evaluate feature performance quantitatively, we use multiple measures. These include standard measures such as Recall-Precision (RP), Average Precision (AP) and Spearman's rank correlation ρ. For the RP evaluation we use images with significant agreement between individuals. Images with a ground truth score $s^* > 0.75$ are taken as positive and $s^* < 0.5$ as negative examples. Images with in-between scores are excluded in the computation of RP, as there is no clear agreement between individuals.

In addition, we use the Top_N score, which quantifies how well the computer-ranked top N images agree with the human ranking. Suppose that s^*_i is the human interestingness score of image I_i, then $Top_N := \frac{\sum_{i \in P_N} s^*_i}{\sum_{i \in S_N} s^*_i}$, where P_N is the set of N images ranked highest by a method, and S_N the set of N images ranked highest according to human consensus. As can be easily seen, $Top_N \in [0, 1]$, where a higher value corresponds to a better performance of the algorithm.

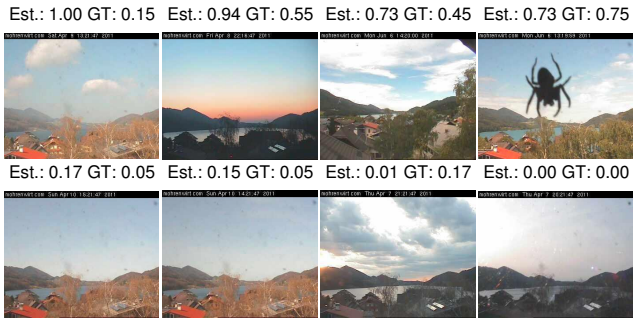
We use the following datasets: Firstly, a set of webcam sequences [12]. Since the presented webcam images are sequential evolving, there is a strong context in which a viewer rates interestingness. Secondly, we use the 8 scene



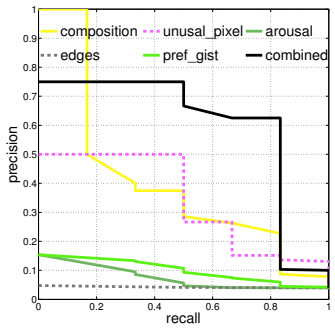
(a) Human labeling. **Top:** most interesting **Bottom:** least interesting.



(b) Interestingness vs. predicted score for Sequence 1.



(c) Predicted interestingness.
Top: most interesting **Bottom:** least interesting.



(d) Recall-Precision curve (Seq. 1). We show the combination along with the five highest weighted cues.

Figure 3: An example (Sequence 1) out of the 20 webcam sequences [12] (**GT:** ground truth, **Est:** the estimated scores from our method).

category dataset [18], which provides some weaker semantic context. Last, we use the memorability dataset [14], which contains arbitrary photographs and offers practically no context. The overview of the results is shown in Tab. 1.

4.1. Strong context: Webcam dataset

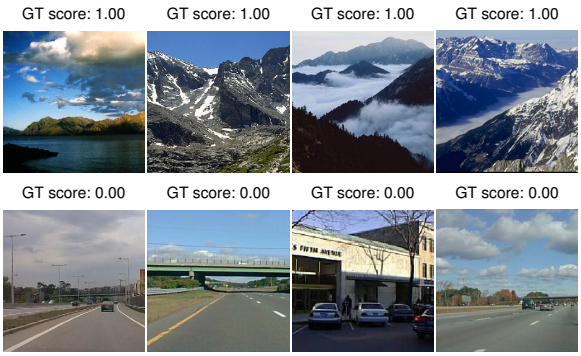
This dataset consists of 20 different webcam streams, with 159 images each. It is annotated with interestingness ground truth, acquired in a psychological study [12]. The interestingness score of an image is calculated as the fraction of people who considered it interesting. There are only a few interesting events in these streams (mean interestingness score of 0.15). Interestingness is highly subjective and there are individuals who did not consider any image interesting in some sequences. An example of these sequences is shown in Fig. 3a.

As a consequence of the low mean interestingness, we use different thresholds for RP calculation: $s^* > 0.5$ as positive and $s^* < 0.25$ as negative samples, which results in about five interesting images per sequence, on average. As we are interested in determining the frames with high interest, (*c.f.* Fig. 3b), Top_5 scores provide a good characterization. We tested each sequence separately and split the remaining sequences randomly into training and validation sets (80% for training / 20% for validation) to train the SVRs and the combination of the features.

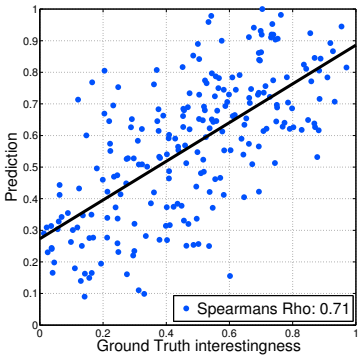
Following the setup of the experiment, we sequentially added the images to the database, as they were tested. *i.e.* the unusualness scores are computed with respect to the previous frames only (while [12] uses the whole sequence).

Results. The mean performance over all 20 sequences is shown in Tab. 1. Results for a sample sequence are shown in Fig. 3. Fig. 3a shows frames of the sample sequence, while Fig. 3c shows the top and bottom predictions of our algorithm. Fig. 3b shows the correlation of predicted interestingness and ground truth score and Fig. 3d plots the Recall-Precision curve for the combination of features along with the five single features having the highest weights.

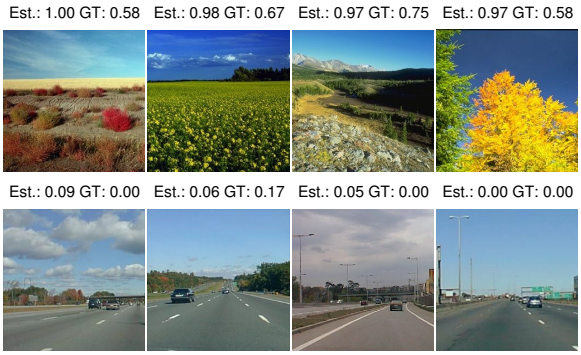
Outlier methods perform best in this setting. Yet, not everything predicted as unusual is rated as interesting by humans, *e.g.* for image 1, Fig. 3c, the method overestimates interestingness, because of cloud formations. This is not unusual at the *semantic* level and therefore not considered interesting by humans. Other typical failure cases include camera shifts (global outlier methods) and direct sunlight or shades. Aesthetics and general preference features show a lower performance. When comparing median scores of our approach to [12] we achieve comparable performance (**AP**: 0.39 (ours) vs. 0.36 [12]; $\mathbf{Top_3}$: 0.66 (ours) vs. 0.72).



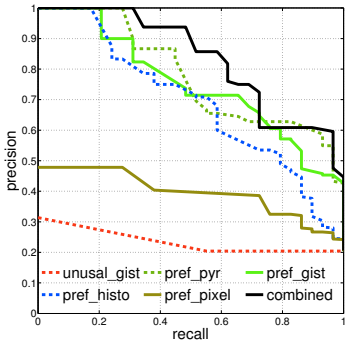
(a) Human labeling. **Top:** most interesting **Bottom:** least interesting.



(b) Interestingness vs. predicted score.



(c) Predicted interestingness.
Top: most interesting **Bottom:** least interesting.



(d) Recall-Precision curve. We show the combination along with the five highest weighted cues.

Figure 4: The 8 scene category dataset (**GT:** ground truth, **Est:** the estimated scores from our method).

4.2. Weak context: Scene categories dataset

The 8 scene categories dataset of Oliva and Torralba [18] consists of 2'688 images with a fixed size of 256×256 pixels. The images are annotated with their scene categories, which allows us to investigate the correlation between scene types and interestingness. The images are typical scenes from one of the 8 categories (coast, mountain, forest, open country, street, inside city, tall buildings and highways). Examples are shown in Fig. 4a.

We extended this dataset with an interestingness score by setting up a simple binary task on Amazon Mechanical Turk. A worker was presented with a pair of randomly selected images from different scene types and asked to choose the one he/she considered more interesting. 20 pairs were grouped into one task. The interestingness score of an image was calculated as the fraction of selections over views. We used this approximation, as a full pairwise comparison ($O(n^2)$) is infeasible for a dataset of this size. Every image was viewed by 11.9 workers on average (min. 10), which is equal to the number of views of [13]. To ensure high quality results, only workers with the "Masters" level were allowed for the task.

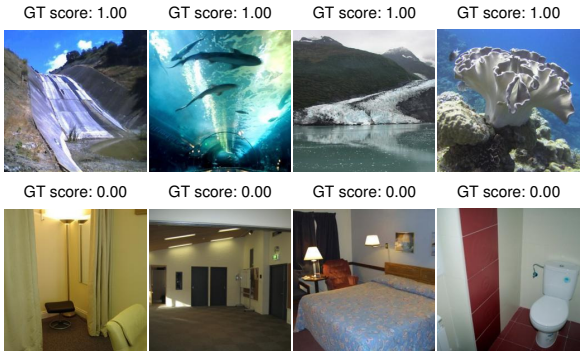
Results. Fig. 4 and Tab. 1 show the results of our features on this dataset. The scene categories provide a weak context, given by the prior on the scene type, which allows to capture novelty/unusualness, as outliers to what are typical images of a certain scene category. Nonetheless, all outlier methods perform less good on this dataset. Novelty is not only harder to capture in this setting, it is also less clearly defined, than in the case of webcams. The algorithm can only capture unusualness with respect to the training images (the prior knowledge of our algorithms), not the observer's prior experience. Furthermore this dataset contains very few unusual images. Therefore a viewer mainly rates the images in this dataset according to aesthetics and general preferences, which transpires from the performance of the individual features.

General preference features yield the highest performance, as they are able to capture scene type and illumination effects (s^{pref}_{hist}), such as the color of a sunset. The features learn the preference for certain scene types (*c.f.* Sec. 2, Fig. 2c) and the aversion for road scenes.

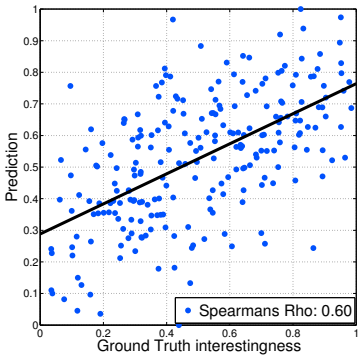
4.3. Arbitrary photos: Memorability dataset

The memorability dataset consists of 2'222 images with a fixed size of 256×256 pixels. It was introduced in [14] by Isola *et al.* and further extended in [13] to investigate the memorability of images (see examples in Fig. 5).

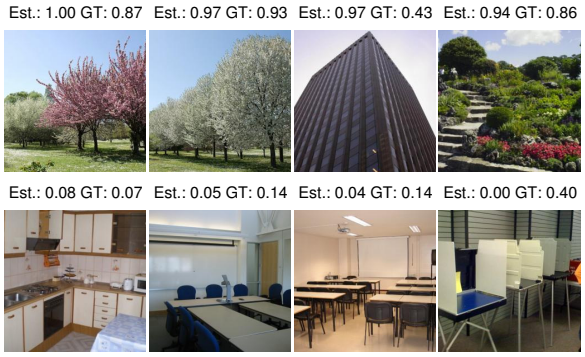
Isola *et al.* [13] included an attribute on interest in



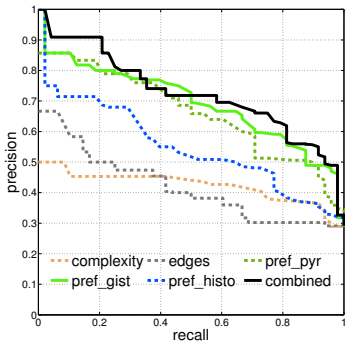
(a) Human labeling. **Top:** most interesting **Bottom:** least interesting.



(b) Interestingness vs. predicted score.



(c) Predicted interestingness.
Top: most interesting **Bottom:** least interesting.



(d) Recall-Precision curve. We show the combination along with the five highest weighted cues.

Figure 5: The memorability dataset (**GT:** ground truth, **Est:** the estimated scores from our method).

their study (attribute "is_interesting"). In their experimental setting, they asked a user to classify an image as interesting/non-interesting. In contrast, we conducted the same study as in Sec. 4.2 on this dataset: We performed a binary experiment, where a user had to select the more interesting image from a pair. The availability of these two experiments allows us to analyze and compare them. Despite the different experimental setting, the scores obtained show a strong correlation ($\rho = 0.63$), suggesting that images hold an intrinsic interestingness.

Results. Fig. 5 and Tab. 1 show the results of our features on this dataset. The trained regressor for general preferences performs best. Not surprisingly, unusualness features perform badly. Based on the psychological findings (*c.f.* Sec. 2) unusualness/novelty probably remains equally important here. Unfortunately, we are not able to capture it for two reasons: (i) What is unusual or novel, in this unconstrained setting, depends on the prior knowledge of the observers, which is unknown to the algorithm. (ii) Semantics are crucial in the appraisal of what is unusual in this dataset. Take, for example, image 3 in the top row of Fig. 5a. This image is interesting, as it shows the end of a glacier. To predict the interestingness of such an image correctly, we need to understand such semantics.

5. Conclusion

Interestingness is an important image property. It is clearly subjective and depends, to a certain degree, on personal preferences and prior knowledge. Nonetheless there exists a substantial agreement about it among observers. This also allowed us to capture it computationally. We proposed a set of features able to capture interestingness in varying contexts. With strong context, such as for static webcams, unusualness is the most important cue for interestingness. In single, context-free images, general preferences for certain scene types are more important. Fig. 6 illustrates the importance of the different interestingness cues as context gets weaker. Unusualness, while remaining important, becomes more difficult to capture with weak contexts. To overcome the current limitations of interestingness prediction, one would need: (i) an extensive knowledge of what is known to most people, (ii) algorithms able to capture unusualness at the semantic level and (iii) knowledge about personal preferences of the observer.

Acknowledgments. We thank Michel Druey and Michael Herzog for their helpful discussions. This work was supported by the European Research Council (ERC) under the project VarCity (#273940) and the Swiss CTI under project no. 15769.1, "Relevance feedback mechanisms for video surveillance".

Data. The used experimental data is available on the authors web-page.

Context	Cue	Feature	ρ	AP	Top_5
Strong Webcams [12] Static camera: 20 different outdoor sequences	Unusual	compose	**0.29**	**0.35**	0.51
		pixel	0.23	0.22	**0.53**
		pyr	0.01	0.10	0.31
		gist	0.03	0.12	0.28
	Aesthetic	arousal	0.13	0.24	0.41
		complex	0.09	0.26	0.48
		colorful	-0.06	0.06	0.26
		edges	-0.04	0.07	0.34
		contrast	0.10	0.15	0.41
	Pref.	pixel	0.04	0.11	0.35
		pyr	0.05	0.10	0.31
		gist	0.16	0.18	0.39
		colorhist	0.05	0.12	0.36
		combined	**0.32**	0.39	0.57
		comb. decorr.	0.31	**0.42**	**0.61**
		chance	0	0.04	0.25
Weak Scene categories [18] 8 scenes types: coast, mountain, forest, open country, street, inside city, tall building, highway	Unusual	compose	0.18	0.28	0.38
		pixel	0.23	0.32	0.32
		pyr	0.17	0.27	0.66
		gist	0.19	0.23	0.47
	Aesthetic	arousal	0.43	0.45	0.65
		complex	0.19	0.31	0.53
		colorful	0.24	0.33	0.67
		edges	0.30	0.34	0.51
		contrast	0.19	0.34	0.62
	Pref.	pixel	0.43	0.40	0.62
		pyr	0.64	**0.78**	0.70
		gist	**0.67**	0.75	0.76
		colorhist	0.54	0.69	**0.83**
		combined	**0.71**	**0.83**	**0.68**
		comb. decorr.	0.70	**0.83**	**0.68**
		chance	0	0.26	0.48
None Memorability[14] Arbitrary photos: Indoor, Outdoor, man-made, natural, people, animals	Unusual	compose	0.10	0.35	0.46
		pixel	0.01	0.31	0.65
		pyr	-0.11	0.29	0.60
		gist	-0.01	0.30	0.45
	Aesthetic	arousal	-0.03	0.31	0.47
		complex	0.27	0.42	0.63
		colorful	0.03	0.34	0.61
		edges	0.11	0.42	0.55
		contrast	0.05	0.33	0.67
	Pref.	pixel	0.25	0.51	0.67
		pyr	0.52	0.66	**0.78**
		gist	**0.58**	**0.69**	0.77
		colorhist	0.33	0.55	0.64
		combined	**0.60**	0.73	**0.82**
		comb. decorr.	**0.60**	**0.77**	0.80
		chance	0	0.26	0.47

Table 1: The interestingness cues and their performance on the 3 datasets. We highlight **best combination** and **best** and second best single feature.

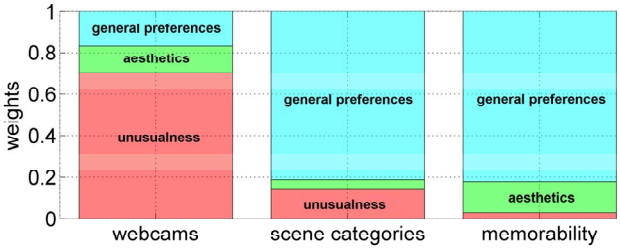


Figure 6: The normalized weights for the feature combinations. The importance of unusualness features decreases, as the context becomes weak.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *PAMI*, 2012.

[2] D. Berlyne. *Conflict, arousal, and curiosity*. McGraw-Hill, 1960.

[3] I. Biederman and E. Vessel. Perceptual Pleasure and the Brain. *American Scientist*, 2006.

[4] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *ICCV*, 2005.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *ACM Sigmod*, 2000.

[7] D. S. Butterfield, C. Fake, C. J. Henderson-Begg, and S. Mourachov. Interestingness ranking of media objects, patent application 20060242139.

[8] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2011.

[9] A. Chen, P. Darst, and R. Pangrazi. An examination of situational interest and its sources. *Brit. J. of Edu. Psychology*, 2001.

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*. 2006.

[11] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.

[12] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in Webcam-Streams. *ACM MM*, 2013.

[13] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. *J. of Vision*, 2012.

[14] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *CVPR*, 2011.

[15] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[17] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. *ACM MM*, 2010.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[19] T. Schaul, L. Pape, T. Glasmachers, V. Graziano, and J. Schmidhuber. Coherence progress: A measure of interestingness based on fixed compressors. In *AGI*, 2011.

[20] P. J. Silvia. Interest - The Curious Emotion. *CDPS*, 2008.

[21] C. A. Smith and P. C. Ellsworth. Patterns of cognitive appraisal in emotion. *J. of of Personality and Social Psychology*, 1985.

[22] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*. 2010.

[23] A. Torralba. How many pixels make an image. *Visual neuroscience*, 2009.

[24] S. A. Turner Jr and P. J. Silvia. Must interesting things be pleasant? A test of competing appraisal structures. *Emotion*, 2006.

[25] P. Valdez and A. Mehrabian. Effects of color on emotions. *J. of Experimental Psychology*, 1994.

[26] E. A. Vessel and N. Rubin. Beauty and the beholder : Highly individual taste for abstract , but not real-world images. *J. of Vision*, 2010.

[27] G. K. Wallace. The JPEG still picture compression standard. *Comm. of the ACM*, 1991.